# Improving Demographic Information for Address Based Sampling (ABS) frames

American Association for Public Opinion Research Annual Conference
New Orleans, LA, May 21st, 2017

Joe McMichael & Jamie Ridenhour

http://artnc.org/works-of-art/new-orleans-ragging-home

# What are most doing now?

- Appending area-level demographics
  - Decennial Census
  - American Community Survey

- Appending address-level demographics
  - Vendors (i.e. MSG)

What if we used predictive models built from previous survey?

# RTI Enhanced ABS Frame

# Enhanced ABS Frame

- CDS (ABS foundation)
- Geocode addresses
- Area-level demographics
  - Census PDB
  - Decennial Census
  - ACS
- Address & person-level
  - Acxiom InfoBase
  - Many sources (black-box)
  - Completeness varies
  - Accuracy varies
- RTI modeled demographics
- City-style, PO Box, etc
- Vacancy status
- Single vs. Multi-family
- Census block group demo
- Child age group
- Adult name
- Adult age
- Adult race
- Income
- Education
- Subscribe to Cat Fancy (joking)
- many others

# Methods

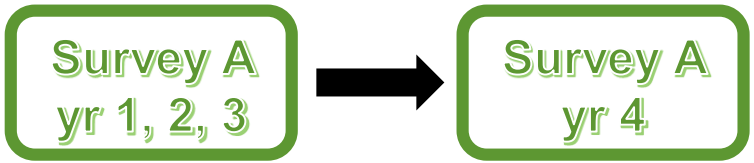
# Survey Data

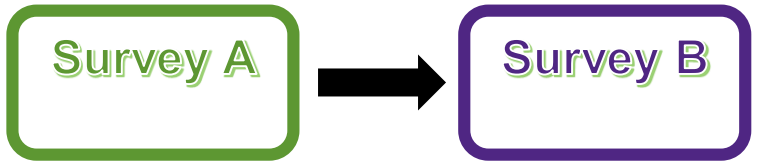
Survey A

- Mail survey
- Single state (a large one)
- Sample - 73,000
- Respondents - 26,500
- 36% return rate
- Collected over four years

Survey B

- Multi-mode national survey
- Consider mail screener only
- Sample – 18,000
- Respondents - 8,000
- 44% return rate



# Inspired by the Census Return Rate Challenge

- Kaggle competition
  - Census Return Rate Challenge
  - Predict mail return rates for the Decennial Census
  - Top models employed machine learning ensemble methods

- Census used most important predictors

# Model building

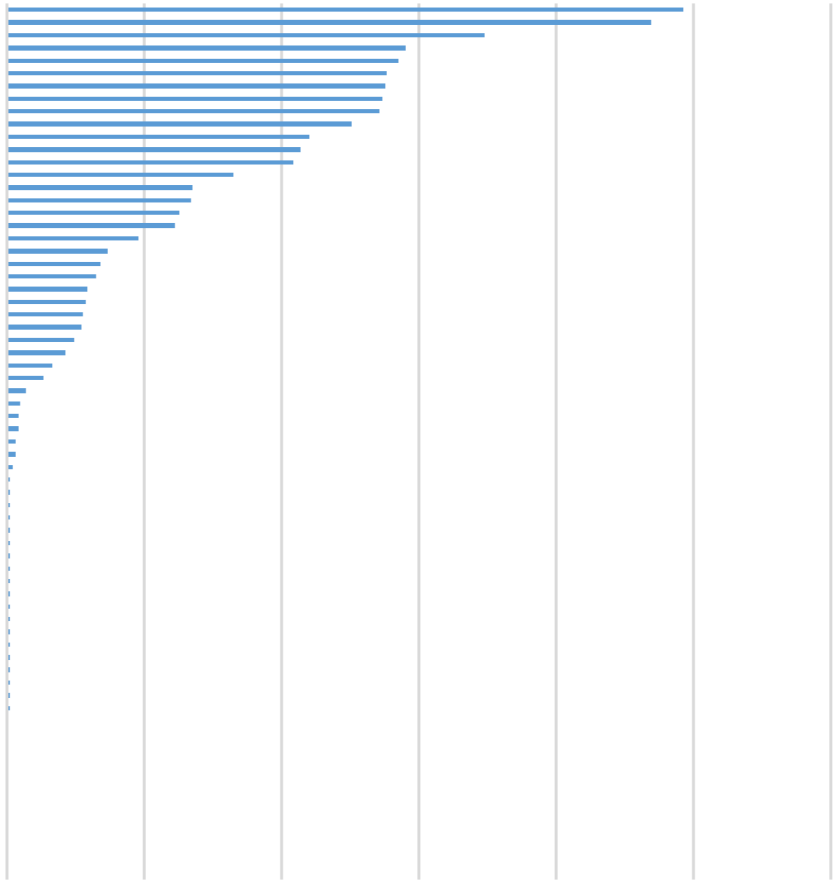
Random forest for variable selection

- Python 3, SciKit, RandomForestClassifier
- 205 predictors
- Variable importance (see next slide for example)

Logistic Regression

- Top predictors from random forest
- Split into training and validation data
- Stepwise variable selection
- ASE of model with validation data
- SAS HPLOGISTIC

# Variable Importance Example



## Highest Ranked Variables

- LRS (PDB)
- DOB age 65+ (Acxiom)

---

- High rise (CDS)
- Vacant (CDS)

---

- % white alone (Decennial Census)
- PersonicX generation categories (Acxiom)
- % College Grad (ACS)
- Has a surname (Acxiom)
- Has DOB (Acxiom)
- Hispanic Surname (Acxiom)

---

- Has child (Acxiom)
- DOB 40-49 (Acxiom)
- DOB 50-59 (Acxiom)
- Black Surname (Acxiom)
- % Other Language (ACS)

# Evaluation

Area under the curve (AUC)

- Measures classification error of predicted vs. observed
- Range 0.5 to 1
- 1 is perfect
- 0.5 is useless

Density Strata

- 6 evenly sized strata
- Based on the modeled predicted probability

# Results

| Density Stratum | Child in HH 13-17 | Hispanic | Education HS or less | Tobacco Use | Adult 18-24 | Adult in HH 18-24 |
|---|---|---|---|---|---|---|
| High - 1 | 34% | 56% | 45% | 29% | 20% | 35% |
| 2 | 18% | 16% | 29% | 25% | 9% | 24% |
| 3 | 14% | 7% | 25% | 18% | 3% | 20% |
| 4 | 9% | 4% | 17% | 20% | 3% | 16% |
| 5 | 3% | 3% | 13% | 14% | 1% | 12% |
| Low - 6 | 1% | 2% | 10% | 11% | 1% | 8% |
| AUC | 0.77 | 0.87 | 0.69 | 0.61 | 0.80 | 0.67 |
| Expected Population Prevalence | 13% | 15% | 40% | 18% | 12% | 15% |
| Observed Average | 13% | 15% | 23% | 19% | 6% | 19% |

# Discussion

- Modeled demographics can improve density stratification
- Sample design
  - Useful for rare populations
  - Optimal sample allocations need accurate predicted prevalence
- Data collection interventions and protocols
- Weight adjustments

# Next steps

- Feature Engineering
- Explore other ensemble methods
- Evaluate additional survey data and outcomes
- Evaluate additional auxiliary data sources

**Joe McMichael**
Research Statistician
919.485.5519
mcmichael@rti.org

Suggested citation:

McMichael, J. P., & Ridenhour, J.L. (2017). **Improving Demographic Information for Address Based Sampling (ABS) frames.** Presented at the American Association for Public Opinion Research (AAPOR) annual conference, New Orleans, LA.